

## ENHANCED SENTIMENTAL ANALYSIS FOR SPEECH SYNTHESIS BASED ON PROSODY FEATURE MODIFICATION USING TD-PSOLA TECHNIQUE FOR ENGLISH

**B. Sudhakar**

Assistant Professor, Department of Electrical Engineering, Annamalai University, Chidambaram,  
Tamil Nadu, India.

---

### ARTICLE INFO

#### **Article History:**

Received: 14 Jun 2015;

Received in revised form:

20 Jun 2015;

Accepted: 22 Jun 2015;

Published online: 30 Jun 2015.

---

#### **Key words:**

Time Domine Pitch Synchronous  
overlap Add (TD-PSOLA),  
Tamil Text to Speech (TTS),  
Sentimental Speech Synthesis  
(SSS).

---

### ABSTRACT

In recent years the synthesis of sentimental speech has various applications in customer services in the area of analyzing business adds in social media, mobile services and human- computer interaction etc. An enhanced sentimental analysis for English Text to Speech (TTS) synthesis systems are produced by modifying prosody feature using Time Domine Pitch Synchronous Over Lap Add (TD-PSOLA) technique has been proposed in this paper. The existing speech synthesis systems have lack of naturalness in output speech and emotions. For experiment analysis four various types of emotions has been produced. The experiment results shows that the naturalness of synthesized output has been enhanced in proposed emotional speech synthesis system to achieves an enhanced performance.

*Copyright © 2015 IJASRD. This is an open access article distributed under the Creative Common Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.*

---

### INTRODUCTION

In recent years ,TTS system has become one of the most important research area due to its important in various applications. Text to speech synthesis is the process of converting normal text to speech signal<sup>[1]</sup>. An important disadvantage existing in the machines could not express with human sentiments Sentiment analysis is the natural language processing task dealing with sentiment detection and classification from texts. The ultimate motive of Sentimental Speech Synthesis (SSS) has to produce the natural speech as human voice. There are two main methods are used for speech production. These methods are formant synthesis and concatenation synthesis is illustrated in<sup>[2]</sup>. The formant synthesizer utilizes a simple model of speech generation and a set of rules to generate the speech. But the formant synthesis is not frequently utilized, because the resulting speech has lack of perfection. Alternatively concatenative synthesis links recordings of human speaker to generate the synthetic speech<sup>[3]</sup>. The generating utterance are more natural. In order to produce diversified emotions (or) sentiments the system needs a large size of data base. To

---

**How to cite this article:** Sudhakar, B., (2015). "Enhanced Sentimental Analysis for Speech Synthesis Based on Prosody Feature Modification Using TD-PSOLA Technique for English". *International Journal of Advanced Scientific Research & Development (IJASRD)*, 02 (02/I), pp. 59 – 64.

solve this problem researchers found prosodic features extraction to enhance the naturalness of sentimental output<sup>[4][5]</sup>. In this aspect very few numbers of speech corpora is needed. Diversified types of sentiments could be liked in to synthesized speech by changing appropriate to acoustic parameters either fundamental frequency or the speech contour duration, and then incorporate the concatenative approaches such as TD-PSOLA (Pitch Synchronization Overlap Add) technique<sup>[6]</sup>. This paper proposed sentimental analysis for TTS synthesis systems are produced by modifying prosody feature using PSOLA technique for English language has been implemented.

## IMPORTANCE OF PROSODY IN TTS SYSTEM

Prosody is one of the key components of speech synthesizers, which allows implementing complex wave of physical, phonetic effects that is being employed to express attitude, assumptions and attention as a parallel channel in our daily speech communication<sup>[8][9]</sup>. In general any communication is a collection of two phases: denotation, which represents the written content or the spoken content and connotation, which represent the emotional and attention effects intended by the speaker or inferred by a listener. Prosody plays an important role to improve the naturalness of emotional speech synthesis system. The speech utterances, which are retrieved from emotional speech database with four diversified types of emotions “Happy”, “Fear”, “Neutral” and “Sad” are split in to a set of units. For the concatenative speech synthesis there are multiple choices for the type of unit. The prominently utilized types include words, syllables, phonemes and diphones<sup>[10]</sup>. The basic synthesis building block is the speech unit. A longer unit length is taken to increase the quality of synthesis.

The pitch, energy and duration are enumerated using prosodic analysis for each segment. Using the estimated parameters the prosody feature templates for diversified emotions are generated.

## PROSODY FEATURES CALCULATION

Among the different prosodic features two features fundamental frequency and energy are taken for calculations<sup>[11]</sup>.

### 3.1 Calculation of Fundamental Frequency

The fundamental frequency  $F_0$  speech signal can be calculated either by time domain (or) frequency domain<sup>[13][14]</sup>. The average value of these two measurements provides the final estimate of  $F_0$ . The fundamental frequency  $F_0$  can be directly calculated through time domain auto-correlation method. The waveform utilizing the auto-correlation function which is expected to show peaks at delays corresponding to multiples of the glottal wave period ( $1/F_0$ ).

The frequency domain cepstrum approach is used to calculate the  $F_0$  indirectly. If the log magnitude spectrum contains many regularly spaced harmonics, then the Fourier analysis of the spectrum will expose a peak corresponding to the spacing between the harmonics: *i.e.*, the fundamental frequency<sup>[12]</sup>.

### 3.2 Calculation of Energy

Speech signal is a non-stationary time varying signal. However, it could be viewed as a stationary signal in a short span ranging between 15 ms and 30ms<sup>[15]</sup>. The short span energy is calculated as

$$E_g = \sum_{p=g-j+1} (s[p]h[g-p]) \dots\dots\dots (1)$$

where s is the speech signal, s[p] is the speech signal, h[g-p] is the applied window g = tT, T represents frame shift and t is the integer.

### PROPOSED TTS SYSTEM USING TD-PSOLA

It is a prominently used concatenative synthesis technique to produce naturalness in the generated speech signal. The ultimate aim of TD-PSOLA technique is to alter the pitch directly on the speech waveform. The TD-PSOLA technique sequentially follows three steps, namely (i) pitch synchronization analysis; (ii) pitch synchronization modification; and (iii) pitch synchronization synthesis. Pitch synchronization analysis plays an important role of TD-PSOLA technique, it executes two tasks: fundamental frequency detection and pitch mark.

Let  $X_n(m)$  denotes the windowed short time signal:

$$X_n(m) = h_n(t_n - m)y(m) \dots\dots\dots (2)$$

where  $t_n$  is the mark point of pitch,  $h_n$  is the window sequence. Pitch synchronization modification links the pitch mark by changing the duration (insert or delete the sequence with the length of pitch duration) and tone (increase or decrease the fundamental frequency). The pitch synchronization synthesis adds the new sequence signal produced in the previous step.

$$X(m) = \frac{\sum b_j x_j(m) h_j(t_j - m)}{\sum h_j^2(t_j - m)} \dots\dots\dots (3)$$

where  $t_j$  is the new pitch mark,  $h_j$  is the synthesized window sequence,  $b_j$  is the weight to compensate the energy loss when modifying the pitch value.

**Figure – 1:** A Flow Diagram of the Emotional Speech Synthesis System using TD-PSOLA

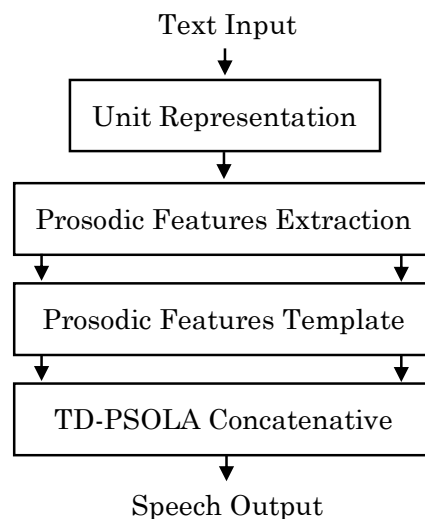
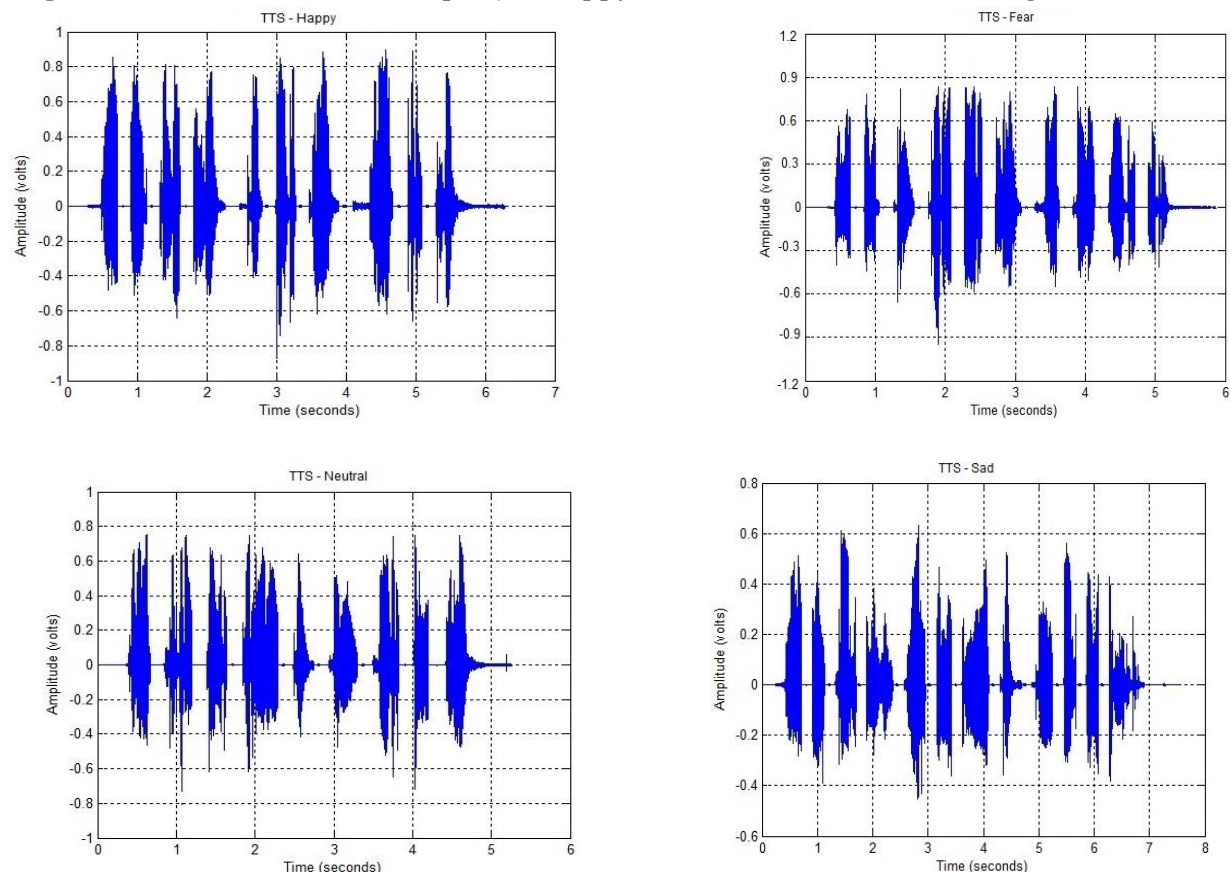


Fig. 1 shows the flow diagram of emotional speech synthesis system using TD-PSOLA technique. English text message with a phoneme string is the input of prosody generator. The duration of each phoneme and the pitch contour is delivered by the prosody generator. Before applying input to the Prosody generator the input is converted into phonemes based on the key strokes involved in the characters present in the input. A database containing the keywords and category of emotion to which it belong. The text is scanned and the keywords present in the text are compared with the contents of the database. The comparison will finalize the value of emotion. The various emotions such as “Happy”, “Fear”, “Neutral” and “Sad” are analyzed by this system. The duration and pitch of each phoneme depends on the content and context of the text. The utterances are segmented into words or units from the input text. Then the prosody features are extracted for each unit and modified according to prosody feature templates with the related emotional types. Finally the synthesized emotional output will be generated through TD-PSOLA concatenative technique, which is used to smooth and modify the boundary unit.

## RESULTS AND DISCUSSION

The simulation results of “Happy”, “Fear”, “Neutral” and “Sad” emotions of the TD-PSOLA based TTS system is shown in Fig. 2. The recorded human voice samples have been taken from noise free environment. It is used as the reference for measure the performance of this TTS system.

**Figure – 2:** *Simulated TTS Output for Happy Fear Neutral Sad Voice using TD-PSOLA*



**Table – 1:** *Variations of Amplitude and Spectral Mismatch of TD-PSOLA based TTS Output for Different Emotions*

<b>Emotions</b>	<b>Amplitude (in volts)</b>	<b>Spectral mismatch (in seconds)</b>
Happy	0.78	0.35
Fear	0.73	0.42
Neutral	0.65	0.48
Sad	0.58	0.54

Table 1 shows the amplitude variations and spectral mismatch variations of the different emotional status for the proposed system. It is inferred that the naturalness has been measured through the amplitude variations and spectral mismatch variations of various emotional sentences with respect to recorded speech. When the amplitude increases and the spectral mismatch decreases the naturalness of the TTS system will be improved.

## CONCLUSION

A novel emotion analysis is designed and implemented using prosodic feature modification method supported by TD-PSOLA technique for English language. The TD-PSOLA technique eliminates the noise produced in the fundamental harmonic component of prosody features. It is also used to smooth and modify the boundary unit to enhance the output of the proposed system in the effective manner. The performance analyses have been implemented for various emotions like “Happy”, “Fear”, “Neutral” and “Sad”. The tabulated performance measures display the highest naturalness achieved by this system for the “Happy” emotion.

## REFERENCES

- [1] Cahn, J. E., (1989) “*Generating Expression in Synthesized Speech*”, Master’s Thesis, MIT. Retrieved from <http://www.media.mit.edu/~cahn/masters-thesis.html>.
- [2] Shirbahadurkar, S. D., Bormane, D. S., & Kazi, R. L., (2010) “*Subjective and Spectrogram Analysis of Speech Synthesizer for Marathi TTS using Concatenative Synthesis*”, Recent Trends in Information, Telecommunication and Computing.
- [3] Bulut, M., Narayan, S., & Syrdal, A., (2002) “*Expressive Speech Synthesis Using a Concatenative Synthesizer*”, Proceedings of ICSLP, 2002, pp. 1265 – 1268.
- [4] Hofer, G., Richmond, K., & Clark, R., “*Informed Blending of Databases for Emotional Speech Synthesis*”, Proceedings of Interspeech 2005, pp. 501-504.
- [5] Schroder, M., (2004) “*Speech and Emotion Research: An Overview of Research Frameworks and a Dimensional Approach to Emotional Speech Synthesis*”, Ph.D. Thesis, Saarland University, Saarland, 2004.
- [6] Rabiner, L. R., & Schafer, R. W., (1978) “*Digital Processing of Speech Signals*”, Prentice-Hall, Inc., Englewood Cliffs, 1978.
- [7] Yang Shun (2010) “*Speech Synthesis Based on PSOLA Algorithm and Modified Pitch Parameters*”. Proceedings of Computational Problem Solving 2010, pp. 296 – 299.

- [8] Jurafsky, D., & Martin, J. H., (2000) “*Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*”. Prentice Hall PTR, Upper Saddle River, NJ, USA, 2000
- [9] Fangzhong Su & Katja Markert (2008) “*From Word to Sense: A Case Study of Subjectivity Recognition*”. In Proceedings of the 22<sup>nd</sup> International Conference on Computational Linguistics, Manchester.
- [10] Vroomen, J., Collier, R., & Mozziconacci, S. J. L., “Duration and Intonation in Emotional Speech”, *Eurospeech 93, Vol. 1*, pp. 577 – 580.
- [11] Burkhardt, F., Paeschke, A., Rolfes, M., et al., (2005) “*A Database of German Emotional Speech*”, Proceedings of Interspeech.
- [12] Sangeetha, S., & Jothilakshmi, S., (2014) “Syllable Based Text to Speech Synthesis System using Auto Associative Neural Network Prosody Prediction”, *International Journal of Speech Technology*, 17 (2), pp. 91 – 98.
- [13] Burkhardt, F., & Sendlmeier, W. F., (2000) “*Verification of Acoustical Correlates of Emotional Speech using Formant-Synthesis*”, ISCA Workshop on Speech & Emotion, Northern Ireland, pp. 151- 156.
- [14] Vroomen, J., Collier, R., & Mozziconacci, S. J. L., (1993) “Duration and Intonation in Emotional Speech”, *Europeech*, 1, pp. 577 – 580.
- [15] He, L., Huang, H., & Lech, M., (2013) “Emotional Speech Synthesis Based on Prosodic Feature Modification”, *Engineering*, 5, pp.73-77, DOI: [10.4236/eng.2013.510B015](https://doi.org/10.4236/eng.2013.510B015).
- [16] Morais, M., & Violaro, F., (2005) “*Data-Driven Text-to-Speech Synthesis*”, XXII Simpósio Brasileiro de Telecomunicações – SBrT’05, 04-08 de Setembro de 2005, Campinas, SP, 1 - 16.
- [17] Kamble, K. S., & Kagalkar, R., (2012) “A Review: Translation of Text to Speech Conversion for Hindi Language”. *International Journal of Science and Research (IJSR)*, 3 (11), pp. 1027 - 1031.